PANDORA - past, present, and future
National web archiving in Australia

Dr Paul Koerbin
Manager Web Archiving
National Library of Australia

National Conference on eResources in Malaysia
Penang, Malaysia, December 2012

National web archiving in Australia

1. PANDORA Web Archive – a brief history
2. Web archiving at the National Library of Australia (NLA) today
3. Issues for future web archiving at the NLA
4. Experiences and lessons learned
5. The importance of web archiving

1. PANDORA - context and history

1. What is web archiving?
2. Why it is important to do it?
3. How did the NLA approach it from the start?
4. Why the NLA has approached web archiving in the manner we have?

- Timeline of major milestones from 1996 to now

What is web archiving?

- Web archiving involves:
  - Selecting or scoping what to collect
  - Collecting content from the web
  - Preserving what we collect
    - Strategies, metadata, maintaining bitstream, actions
  - Providing access to the collection
    - Long term and current
- Creating heritage artefacts
    - Creating the time dimension for the web

What are we collecting?

- Web sites (and all they contain)
  - Complex objects
    - Text, images, media, style elements, client side scripts
    - No control over formats, systems, creation of content
- Includes sites with embedded media
  - lots of formats (mpeg, flv, QT, wmv, rm, Shockwave)
- Content is harvested with crawl robots
  - A browser view not underlying database
  - Dynamic content becomes static HTML

Why do web archiving?

- Statutory responsibility
  - National Library Act (1960)
    - Maintain, develop comprehensive collection relating to Australia and Australian people
    - Make national collection available in the national interest
- Just another publishing medium
  - Prior experience with non-print formats and ephemera
- National leadership
- Vision – recognising importance of new way information is communicated

How did we approach it?

- Selective approach
- Proof of concept approach
- Workflows and infrastructure developed and implemented in an ongoing way
- Initiated collaboration with other collecting institutions

Why did we approach it this way?

- Scalable to available resources
- Staged approach – do what we can
- Work within the legal constraints
- Able to realise objective of current access
- Heuristic – this was pioneering times and options were limited

Timeline

- April 1996: ‘Electronic Unit’ established
  - Part of ‘Acquisitions Branch’
  - 3 staff, 6 months to develop selection (scope) guidelines and identify resources
- September 1996: ‘Australian Serials and Electronic Unit’ established
  - Technical services restructure, multi-tasking, matrix management
  - October 1996 first titles harvested
- November 1996: ‘PANDORA’ born as ‘proof of concept project’
  - **P**reserving and **A**ccessing **N**etworked **Do**cumentary **R**esources of **A**ustralia
  - As at June 1997, 30 titles harvested
- May 1998: public access to PANDORA titles

Timeline (continued)

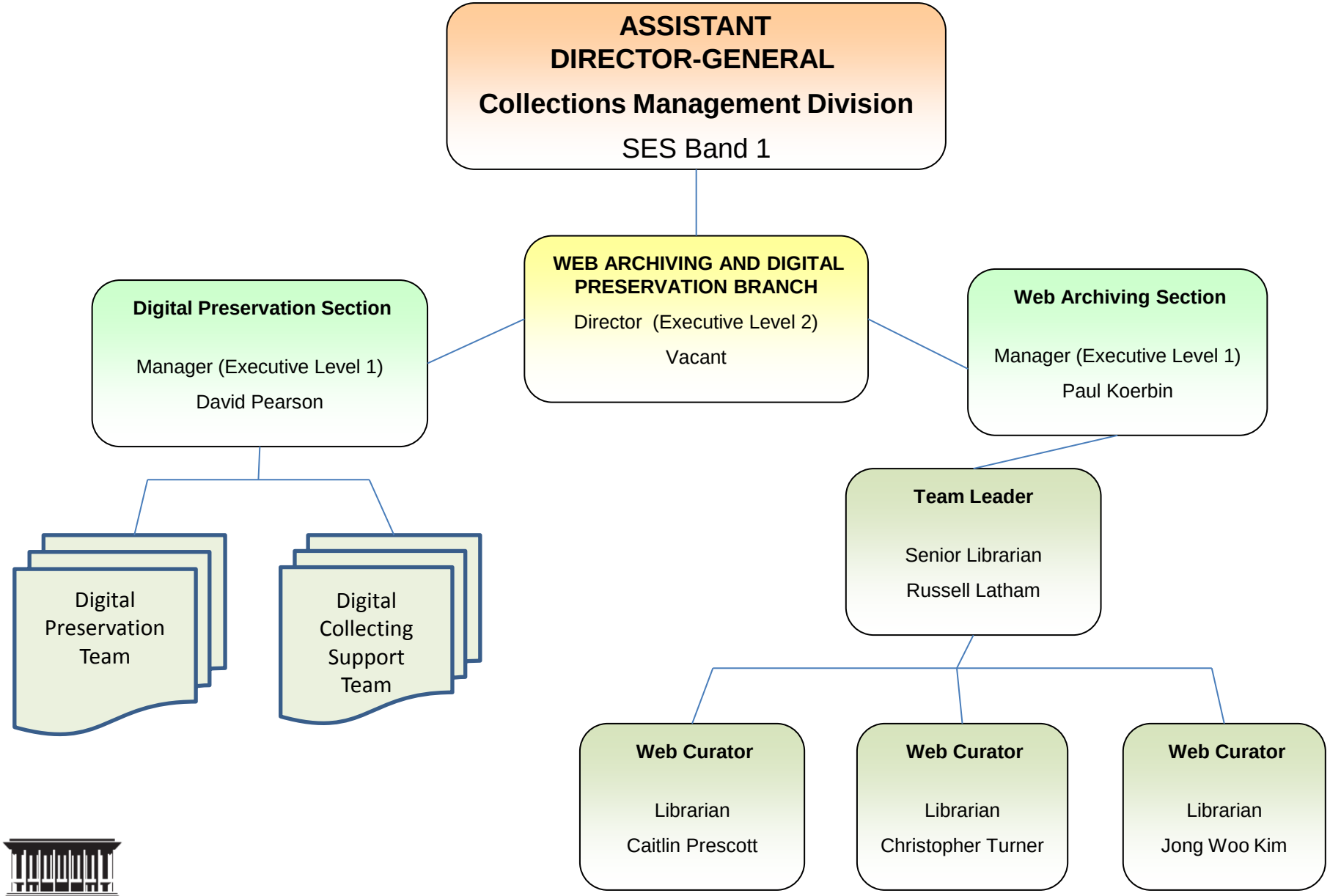
- July 1998: first PANDORA 'partner' began participation
  - 11th participant joined in 2010
- June 2001: PANDAS v.1 released
  - Web archiving workflow system developed by NLA
- 2002: Digital Archiving Branch
  - Our own identity at last!
  - Began first trial of 'mainstreaming' web archiving in Serials and Govt Deposit sections
- August 2002: PANDAS v.2 released
- July 2003: joined IIPC
  - International Internet Preservation Consortium

Timeline (continued)

- 2004: PANDORA added to UNESCO *Australian Memory of the World Register*
- July 2005: first .au domain harvest
  - Subsequent harvests in 2006, 2007, 2008, 2009, 2011, 2012
- Dec. 2006: “Web Archiving and Digital Preservation Branch”
- July 2007: PANDAS v.3 released
- 2010: PANDORA search moved to Trove
- May 2010: Whole-of-government (federal) permission
- March 2011: first gov.au bulk seed list collection
- Dec. 2012: Australian Government Web Archive

2. Web archiving at NLA today

- Organisation
- Staffing
- Collaboration participants
- NLA web archive collections
- Workflows
- PANDAS


ASSISTANT DIRECTOR-GENERAL
Collections Management Division
SES Band 1
WEB ARCHIVING AND DIGITAL PRESERVATION BRANCH
Director (Executive Level 2)
Vacant
Digital Preservation Section
Manager (Executive Level 1)
David Pearson
Web Archiving Section
Manager (Executive Level 1)
Paul Koerbin
Digital Preservation Team
Digital Collecting Support Team
Team Leader
Senior Librarian
Russell Latham
Web Curator
Librarian
Caitlin Prescott
Web Curator
Librarian
Christopher Turner
Web Curator
Librarian
Jong Woo Kim

Staffing – skills and expertise

- Collection development
  - Selection expertise in ‘new media’
  - Corporate objectives, priorities, resources
- Collection management
  - Cataloguing: MARC, LCSH, Dewey
  - PANDORA subjects
- Technical skills
  - Scoping gather filters and settings
  - Harvesting and code problem analysis and resolution (HTML, JavaScript, stylesheets)
  - Understanding web technologies
- Experience and self-learning
  - New technologies, Web 2.0, timely collecting, always new challenges

PANDORA participants

- 11 participants including the NLA
- State and territory libraries (not Tasmania and ACT)
- Major heritage institutions
  - National Film and Sound Archive
  - Australian War Memorial
  - Australian Institute of Aboriginal and Torres Strait Islander Studies
  - National Gallery of Australia

PANDORA participants

- Memorandum of Understanding
  - Respective obligations (NLA and Agencies)
  - Adherence to policy and procedures
- Curatorial and collection management (operational staff)
  - Selection (participants have their own guidelines)
  - Permissions
  - Harvesting – scoping and quality checking
  - Cataloguing
  - Publishing – access through PANDORA

Web archive collections

- Three collecting approaches and collections
  - 'PANDORA Archive' collection
    - Selective web archiving since 1996
  - 'Australian Web Domain' collection
    - Large scale, outsourced (IA), annual collection, since 2005
  - Australian Government Web Archive' collection
    - Bulk seed list harvesting, outsourced (IA), annual collection, since 2011

  Plus Asia-Pacific web collections using the *Archive-It* collecting and hosting service (since 2007)

PANDORA | gov.au 2011-2012 | AusCrawl 2005-2012

Selective

'targets', 'titles'

Small scale

Reactive
Timely
Scheduled

High curation

High access
Controlled

Themed

Curated seed lists
or 'titles'

Moderate scale

Scheduled
Timely

High curation

High access
Controlled

2nd L Domain

e.g. *gov.au*

Moderate to large scale

Scheduled
(moderate control)

Moderate curation

Moderate access
Moderate risk

TL Domain

i.e. *.au*

Large scale

Scheduled
(low control)

Low curation

Low/mod access
High risk

Whole Web

Internet Archive

Large scale

Ongoing
Unscheduled

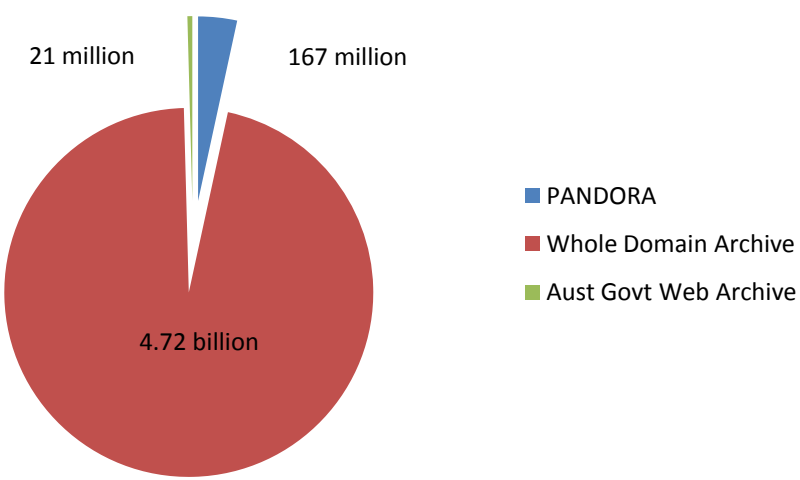
No curation control

No access control

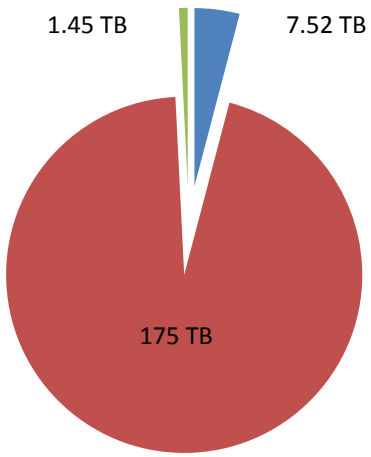
Statistics

	PANDORA Web Archive 1996 – Nov. 2012	Australia Whole Domain Web Archive 2005-2012	Australian Government Web Archive 2011-2012	All collections
Files	167 million	4.72 billion	21 million	4.9 billion
Data	7.52 TB	175 TB	1.45 TB	184 TB

Statistics - all web collections

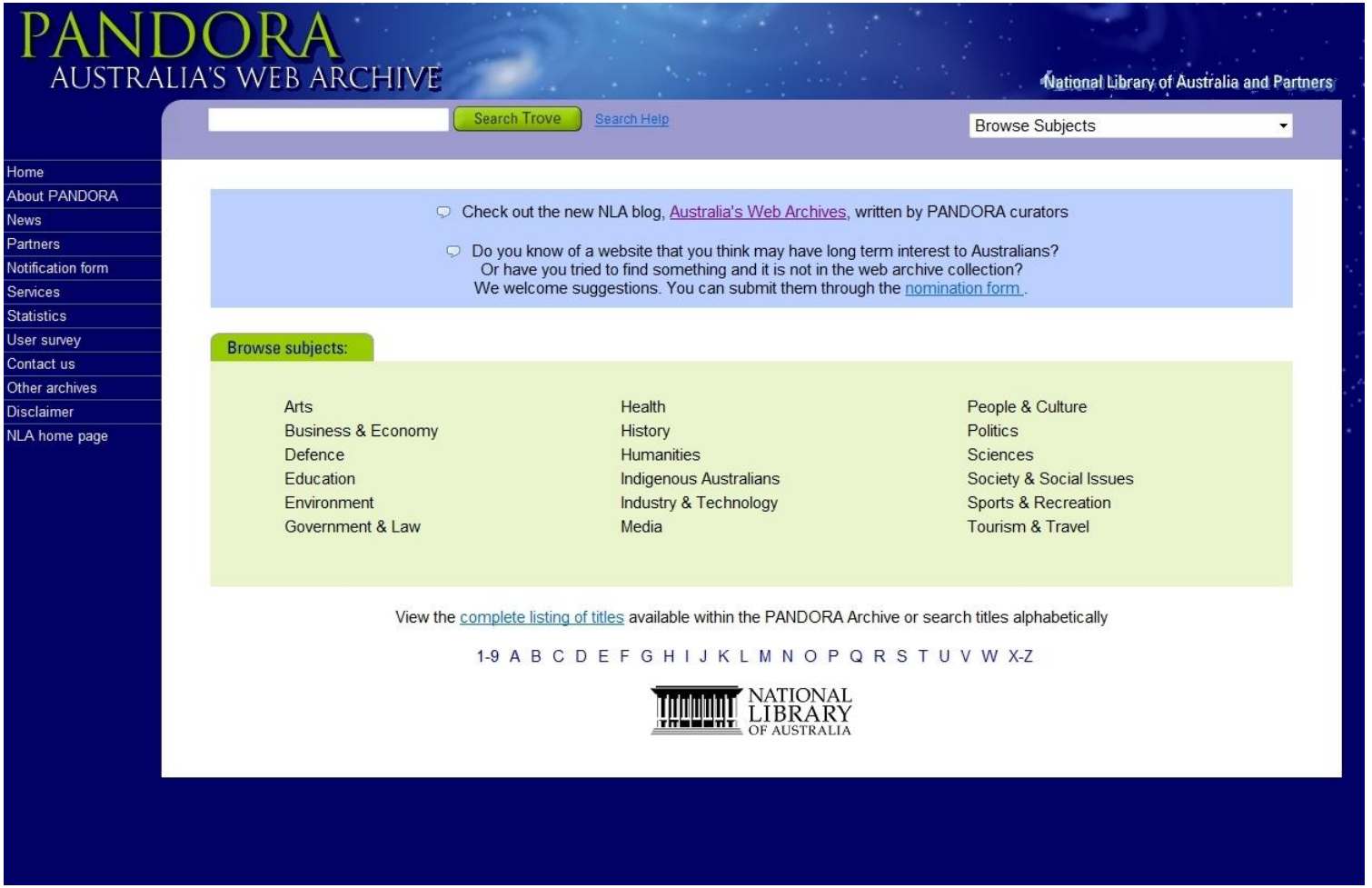


Files archived



Data archived

Web archiving workflow



PANDAS
PANDORA DIGITAL ARCHIVING SYSTEM

Help | Logged in as **pkoerbin** | Access level **Pandas Administrator** | Logout

Search Title

- Home

Main
- Search
- Add Title
- Add Publisher
- Add Indexer
- Add Collection

Administration
- Edit My Details
- Browse Agencies
- Add Agency
- Browse Users
- Add User
- Browse Subjects
- Add Subject
- Browse Restriction Groups
- Add Restriction Group

Tools
- Gather Queue
- Form Letters
- Reports

External Links
- PANDORA
- QA System

Worktrays

National Library of Australia

Expand All | Collapse All

Transferred — Refresh
- View Unacknowledged Titles Transferred to Me [1]

Selection — Refresh
- View Nominated Titles [32]
- View Monitored Titles [52]
- View Selected Titles [50]

Permission — Refresh
- View Permission Requested Titles [158]
- View Permission Granted Titles [32]

Gather — Refresh
- View Scheduled Titles [218]
- View Gathering Titles [2]

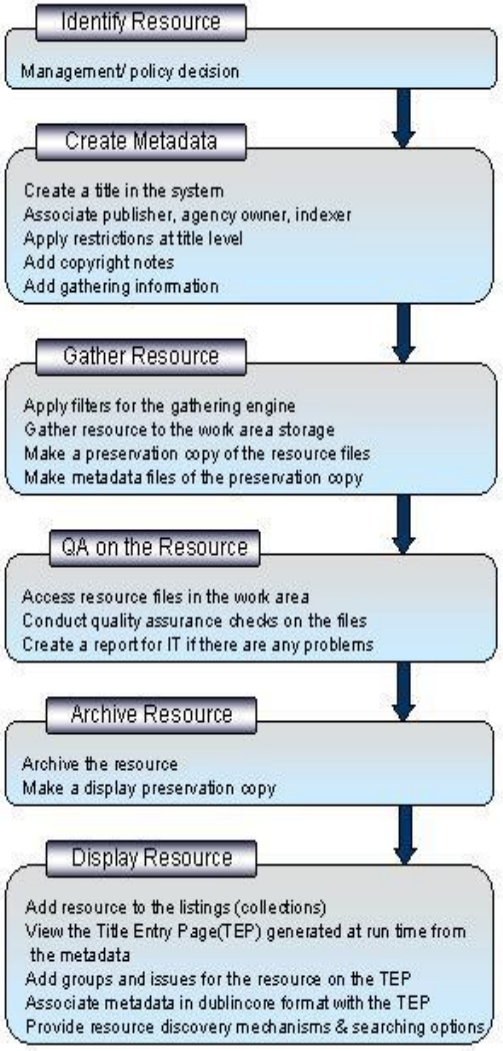
Preserve — Refresh
- View Instances For Upload [12]
- View Instances in QA [17]
- View Instances Referred to IT [1]

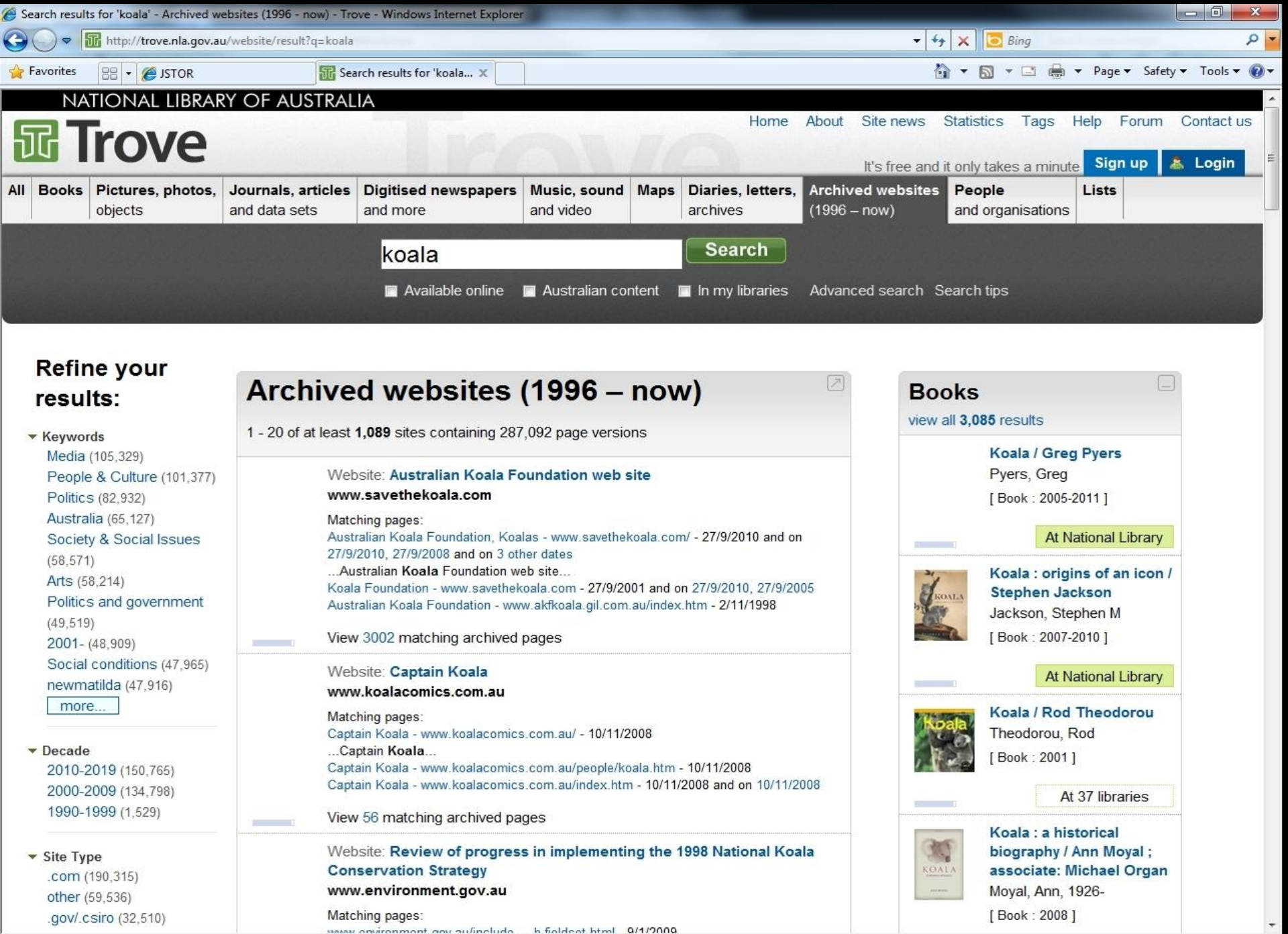
Publish — Refresh
- View Archived Titles [16]

Catalogue — Refresh
- View Titles Requiring Cataloguing [119]

Reports — Refresh
- View Report Queue [2]
- View Generated Reports [30]

PANDAS system workflow diagram




Search results for 'koala' - Archived websites (1996 - now) - Trove - Windows Internet Explorer
http://trove.nla.gov.au/website/result?q=koala
Bing
Favorites
JSTOR
Search results for 'koala...
Page
Safety
Tools
NATIONAL LIBRARY OF AUSTRALIA
Trove
Home
About
Site news
Statistics
Tags
Help
Forum
Contact us
It's free and it only takes a minute
Sign up
Login
All
Books
Pictures, photos, objects
Journals, articles and data sets
Digitised newspapers and more
Music, sound and video
Maps
Diaries, letters, archives
Archived websites (1996 – now)
People and organisations
Lists
koala
Search
Available online
Australian content
In my libraries
Advanced search
Search tips
Refine your results:
Keywords
Media (105,329)
People & Culture (101,377)
Politics (82,932)
Australia (65,127)
Society & Social Issues (58,571)
Arts (58,214)
Politics and government (49,519)
2001- (48,909)
Social conditions (47,965)
newmatilda (47,916)
more...
Decade
2010-2019 (150,765)
2000-2009 (134,798)
1990-1999 (1,529)
Site Type
.com (190,315)
other (59,536)
.gov/.csiro (32,510)
Archived websites (1996 – now)
1 - 20 of at least 1,089 sites containing 287,092 page versions
Website: Australian Koala Foundation web site
www.savethekoala.com
Matching pages:
Australian Koala Foundation, Koalas - www.savethekoala.com/ - 27/9/2010 and on 27/9/2010, 27/9/2008 and on 3 other dates
...Australian Koala Foundation web site...
Koala Foundation - www.savethekoala.com - 27/9/2001 and on 27/9/2010, 27/9/2005
Australian Koala Foundation - www.akfkoala.gil.com.au/index.htm - 2/11/1998
View 3002 matching archived pages
Website: Captain Koala
www.koalacomics.com.au
Matching pages:
Captain Koala - www.koalacomics.com.au/ - 10/11/2008
...Captain Koala...
Captain Koala - www.koalacomics.com.au/people/koala.htm - 10/11/2008
Captain Koala - www.koalacomics.com.au/index.htm - 10/11/2008 and on 10/11/2008
View 56 matching archived pages
Website: Review of progress in implementing the 1998 National Koala Conservation Strategy
www.environment.gov.au
Matching pages:
Books
view all 3,085 results
Koala / Greg Pyers
Pyers, Greg
[Book : 2005-2011]
At National Library
Koala : origins of an icon / Stephen Jackson
Jackson, Stephen M
[Book : 2007-2010]
At National Library
Koala / Rod Theodorou
Theodorou, Rod
[Book : 2001]
At 37 libraries
Koala : a historical biography / Ann Moyal ; associate: Michael Organ
Moyal, Ann, 1926-
[Book : 2008]

3. Issues for the future

- The environment for web archiving
- Organisation, infrastructure and workflows
- Access, search, discovery and promotions

Future web archiving at the NLA

- Issues – the environment
  - Ever increasing scale of information online
  - Dynamic delivery of content
  - Complexity of content
    - Technical and intellectual
  - Changing technology
  - Changing understanding of publishing
  - Legal deposit (collecting) and access rights

Future web archiving at the NLA

- Issues – the organisation
  - Infrastructure
  - DLIR
  - Emphasis on digital collecting
  - New approaches to collaboration
  - New workflows and systems

Future web archiving at the NLA

- Access, search, discovery, promotion
  - Full-text search
  - URL search
  - Browse paths and collections
  - Trove – one search service
  - More innovative discovery and visualisation
  - Research (big) data

4. Experience suggests …

1) Important to take actions as soon as practicable
   - Web content is ephemeral like no other
   - The timing of collecting content is critical
   - Things will change but not get easier or simpler to deal with
   - Do what you can – can't solve all problems up front
   - Gain experience by doing
   - Be as agile in your approach as possible so as to deal with the dynamic nature of the task

Experience suggests …

2) Make it sustainable
   - Understand it is commitment over the long term
     - The preservation objective
   - Consider the implications of your technology and infrastructure choices
   - Chose the right approach for your business
     - in-house or out-sourced service?
   - Engage and support skilled, interested, self-motivated staff

Experience suggests …

3) Focus on purpose and outcomes

- Add value through a focus on access and discovery
- Demonstrate the value of collecting web materials
  - Embrace the responsibility of creating the historic cultural artefacts out of the ephemeral
  - Promote the activity – papers, conferences, social media, website
- Engage stakeholders through collaboration and opportunities to contribute
  - Provided this genuinely advances the objectives

Importance of web archiving

- Major task is to advocate the importance of web archiving
  - Hard to demonstrate when content remains on live site
  - Value of content is variable
  - So much a part of everyday life – easy to overlook value

Importance of web archiving

- Only evidence of our cultural expression from the web is in the archives
- Vulnerable medium – ‘unpublishing’
- Democratic medium – value in understanding our society
- Grey literature – moving to online only
- Avoid the digital ‘black hole’

Thank you.

Dr Paul Koerbin
Manager Web Archiving
National Library of Australia
pkoerbin@nla.gov.au

NATIONAL LIBRARY OF AUSTRALIA nla.gov.au